PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20289
doi: 10.3897/tdwgproceedings.1.20289

OPEN ACCESS

Conference Abstract

Traits in a graph

Jennifer Hammock‡, Katja S Schulz‡

‡ Smithsonian Institution, Washington, United States of America

Corresponding author: Jennifer Hammock (hammockj@si.edu)

Received: 14 Aug 2017 | Published: 14 Aug 2017

Citation: Hammock J, Schulz K (2017) Traits in a graph. Proceedings of TDWG 1: e20289.
https://doi.org/10.3897/tdwgproceedings.1.20289

Abstract

Biodiversity data are well-indexed by taxonomic names. While names reconciliation remains a challenge, there has been tremendous progress in recent years, and integration with available phylogenetic information can support sophisticated analyses for evolutionary questions. However, organisms are also linked to each other by relationships of ecology, geographic proximity, shared habitat, management categories, and other attributes, not yet recorded in a well-structured way.

These data are best modeled as a graph, which makes these relationships explicit, and available for reasoning across - just like taxonomic relationships. This would support broad analyses of life on Earth not only from an evolutionary perspective but also across many other axes.

This case study will describe how several categories of data are being modeled in the Encyclopedia of Life (EOL) v3 using ontology terms. It will focus on several areas where we anticipate sufficient taxonomic coverage to underlie significant search and analytical power: habitat, distribution, body size and metabolism, and provenance.

Habitat and distribution terms are good examples of data terms in well structured hierarchies that could support powerful search. Habitat terms are available from and hierarchically organized in the Environment Ontology (ENVO). Geographic distribution knowledge can often be structured by geographic terms based on verbatim locality text when geocoordinates are not available. Geographic terms are available from several providers, notably Geonames (geonames.org), Marineregions.org and Wikidata. Both

© Hammock J, Schulz K. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

habitat and distribution terms can also be connected to simpler and less formalized but commonly used hierarchies like the World Wildlife Fund (WWF) Ecoregions. The hierarchy information made available for habitat and geography by the semantic structure of these ontologies supports searches like "wetland plants of South America," which requires the intersection of taxonomic, geographic, and habitat hierarchies.

Body size and metabolism traits interact in a particular use case, illustrating the importance of precision of categorical data terms for informing calculations of quantitative traits. The use case EOL is currently working to support is the parameterization of food web interactions in ecological modelling software. Default or starting values are needed for the content of energy (or carbon) within an organism, and the rate of loss thereof through metabolism. This, plus assimilation efficiency, allows the modeling of carbon flow through the food web. Traits available for estimating carbon content and metabolic rate include various measures of body size, for which conversion factors and formulae are available. For phytoplankton, for instance, size may be reported as cell dimensions, cell volume, cell wet mass, cell dry mass, and/or carbon biomass. For an automated tool to derive parameters from these which are fit for use, the different types of data must all be findable, but the measurement types must be distinguished from one another so the correct conversions are performed for each - all in a machine readable way, so the process can be automated. The need for semantically structured data terms in this case is different, but just as critical to the success of the use case.

Future work: Other important structured connections can be made through provenance metadata. These connect taxa and specimens to literature, authors, collectors, wildlife observers and other agents. The Social Media of biodiversity data, rendered explicit, could increase connectivity and communication in the global community - particularly benefitting young researchers in isolated regions without the benefit of professional travel or literature subscriptions. To accomplish this, we must leverage human identifiers such as those made available by Open Researcher And Contributor ID (ORCID) and Wikidata.

Keywords

Graph Data; Terms relationships; Metadata; Search; neo4j

Presenting author

Jennifer Hammock

Acknowledgements

This work is funded by

- the Sloan Foundation

- the Smithsonian Institution
- the National Science Foundation, Award #1636859- BD Spokes: Spoke: South: Collaborative: Using Big Data for Environmental Sustainability: Big Data + AI Technology = Accessible, Usable, Useful Data!